HMM Based Part of Speech Tagging for Hadith Isnad

Abdelkarim Abdelkader
aaabdulkader@uqu.edu.sa
Computer Engineering Department, Faculty of Computing at Algunfdha, Umm Al-Qura University, KSA

Summary
The Hadith is the second source of Islamic jurisprudence after Qur'an. Both sources are indispensable for muslims to practice Islam. All Ahadith are collected and are written. But most books of Hadith contain Ahadith that can be weak or rejected. So, quite a long time, scholars of Hadith have defined laws, rules and principles of Hadith to know the correct Hadith (Sahih) from the fair (Hassen) and weak (Dhaif). Unfortunately, the application of these rules, laws and principles is done manually by the specialists or students until now. The work presented in this paper is part of the automatic treatment of Hadith, and more specifically, it aims to automatically process the chain of narrators (Hadith Isnad) to find its different components and affect for each component its own tag using a statistical method: the Hidden Markov Models (HMM). This method is a power abstraction for times series data and a robust tool for representing probability distributions over sequences of observations. In this paper, we describe an important tool in the Hadith isnad processing: A chunker with HMM. The role of this tool is to decompose the chain of narrators (Isnad) and determine the tag of each part of Isnad (POI). First, we have compiled a tagset containing 13 tags. Then, we have used these tags to manually conceive a corpus of 100 chains of narrators from "Sahih Alboukhari" and we have extracted a lexicon from this corpus. This lexicon is a set of XML documents based on HPSG features and it contains the information of 134 narrators. After that, we have designed and implemented an analyzer based on HMM that permit to assign for each part of Isnad its proper tag and for each narrator its features. The system was tested on 2661 not duplicated Isnad from "Sahih Alboukhari". The obtained result achieved F-scores of 93%.

Keywords:
Hadith; Isnad; HMM; Tagger; POI; Narrator; Segmentation

1. Introduction

To implement NLP (Automatic Language Processing) tools in Arabic, researchers may need:

- Basic modules for sentence and word segmentation, morphological, syntactic and even semantic analysis;
- Language resources (dictionaries, corpora, lexical databases, etc.);
- Resources and comparison modules for the evaluation;
- Language processing utilities (text search tools, statistical tools on annotated texts and corpora, etc.);

Among the necessary and basic modules, morph syntactic tagging is an essential step for carrying out most applications in natural language processing because it makes it possible to identify the lexical and grammatical category to which the words of the text belong. Thus, taggers are an essential module in consumer applications such as automatic grammatical correction, automatic generation of summaries and information retrieval. In general, morph syntactic labeling is a preliminary step that is difficult to avoid in most Arabic NLP applications. It is also very useful in hadith processing to classify or judge isnad.

Despite the huge scientific improvement and computer revolution nowadays, the computerization of Islamic studies in general and in Hadith for specific is still limited, and most of the available applications and encyclopedias are restricted to provide the Hadith in different ways from printed books.

From this point emerged the idea of creating a laboratory for applied researches and applications in the faculty of computer science in Al-Qunfudah to serve Islamic studies and Arabic language, to meet the need of student and researcher in Islamic studies and biography of the Prophet Mohammad Peace Be Upon Him (PBUH), and to facilitate the Hadith studies and the Prophet news and help to know the narrators [1].

This project aims to process Isnad Hadith based on Hidden Markov Model, and to create dynamic ontology kernel which supports researcher to investigate Isnad Hadith and determinate levels of narrators and their relationships. To achieve the planned aim we will perform the following steps:

Prepare database for Hadith narrators which can be used as knowledge base kernel to computerize Islamic studies based on XML technology, XML is considered as the leading standard way for easy prototyping and exchanging information between different systems and applications. Creating this database depends on existing narrator's books and books of "Jarh" and "Ta'adel" [2] and [3].

Manuscript received March 5, 2023
Manuscript revised March 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.3.16

Applying smart algorithm to analyze Isnad Hadith in morphological and syntactical ways using Hidden Markov Model, which is the most applied in natural language processing, to extract narrator and explain how he has received Hadith from his old narrator and to create Isnad tree for all possible paths for each Hadith.
Trace Isnad tree for each narrator, and highlight the correct paths for Isnad, and alert about weak path.
Enrich this database automatically depending on natural language processing techniques and text mining methods.
We have finished the first step and the results are published in [2] [4], [5] and [6]. The work presented in this paper concern the second and third steps. Second section concerns the state of the art. The first par provides a short presentation of Arabic language, Hadith and Isnad. The second part is devoted to show the approach of part of speech tagging of the Arabic language in general and the main previous related works in this field concerning the Hadith segmentation and processing. After that, third section presents the methodology followed to apply the HMM for part of Isnad tagging (POI). The implementation and experimental results are discussed in fourth section. The last section sums up our contributions and outlines some possible future works

2. State of the Art

2.1 An Overview of Arabic Language, Hadith and Isnad

Like other Semitic languages, the Arabic language has a rich and complex morphology. Arabic is written from right to left. The letters are linked together as in the cursive writing English. We use the same punctuation marks as in French, but we usually write them upside down. Capital letters do not exist. The Arabic alphabet is made up of twenty eight letters. The majority of these letters change shape depending on whether they are isolated or written at the beginning, in the middle or at the end of a word. Specificity concerns the optional use of vowels. Vowels are added above or below the letters, in the form of diacritics. They are useful for reading and correctly understanding a text, because they make possible to distinguish words with the same graphic representation. They are useful, in particular, to perform the correct grammatical interpretation of a word regardless of its position in the sentence [7].
The traditional grammar of the Arabic language includes two categories of rules, morphology (الصرف) and syntax (النحو) divided into three areas namely:

- Inflectional morphology which deals with variations in morph-phonological forms and not variations in meaning.
- Derivational morphology, with the root-and-scheme model, where the root provides a general abstract meaning and the scheme assigns the grammatical category simultaneously with functional and semantic features.
- Syntax and syntactic functions such as subject, direct or indirect object, noun complement, etc., in order to determine the correct declension suffixes [8].

The written Arabic language comes in two main forms: Modern Standard Arabic and Classical Arabic. By classical Arabic, we mean the Arabic of the Quran and Islamic sciences (Science of Hadith, jurisprudence, etc.) , and that spoken and written between the first century of the Hegira (corresponding to the 6th century according to the Gregorian calendar) and the 7th century of the Hegira (or the beginning of the 13th century according to the Gregorian calendar). Modern Arabic, on the other hand, corresponds to that which has been spoken and written since the end of the 18th century until today. Now, it is used as the only language of official written communication by speakers and organizations in all Arabic countries. It is also the language of all official oral communications: political speeches, official press releases, television news or broadcast on the radio, scientific communications, etc. It is also used at university and in scientific conferences.

The science of Hadith is a branch of Islamic sciences. This science studies all that has been reported about the prophet of Islam, his words, his deeds and all that approved. It also studies her biography before and after the revelation. The Muslims made a great effort to memorize and transmit everything they heard said and saw done, down to the small details of the signs that the Prophet made with his hands and face. Among the companions of the prophet, there were some who wrote these Hadiths so as not to forget them, but none of them collected all the Hadiths. After the death of the prophet, his companions took on the responsibility of keeping and transmitting the Hadiths to Muslims. Two essential concerns founded this science:

- Preserve the Hadiths to transmit them to all Muslims.
- Prevent any modification or change of the disclosure.

To achieve these two goals, a scientific movement emerged and developed from generation to generation.
The science of hadith is divided into two different branches; the first is the science of reporting the words, deeds and everything that has approved, as well as the detailed description of one's behavior and physique. The second branch consists of studying the rules and provisions

to be respected, the biography of the reporters, the analysis of the Hadiths and their classification.

Each Hadith consists of two parts: Isnad (إسناد) which is the narrator's chain, and Matn (متن) which is the narration itself [5].

Hadith Scholars construct a science called "Mustalah Al Hadith" (Hadith Terminology), that sets the rules of cite, review and determine the layers and generations of narrators. Hadith scholars have a set of rules to differentiate between Sahih (sound), Da'eif (weak), Mutassil (continuous), Munqati' (broken), Marfu' (elevated), and Mawquf (related to narrations of companion) [4].

The books of Mustalah (terminology) talk about a number of classes of hadith according to their status. The following major classifications can be established [9]:

1) According to the reference to a particular authority (Fig.1). The Isnad can stop at the Prophet, the Hadith in this case is called elevated, or it can stop at the Companion, the Hadith in this case is called stopped, or it can stop at the successor and the Hadith is called severed. In all cases the Isnad of Hadith can contain a set of successors and narrators (Fig. 2).

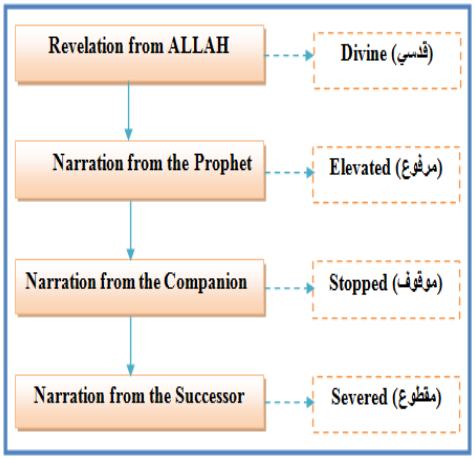

Fig. 1 Classification of Hadith according to the reference to a particular authority.

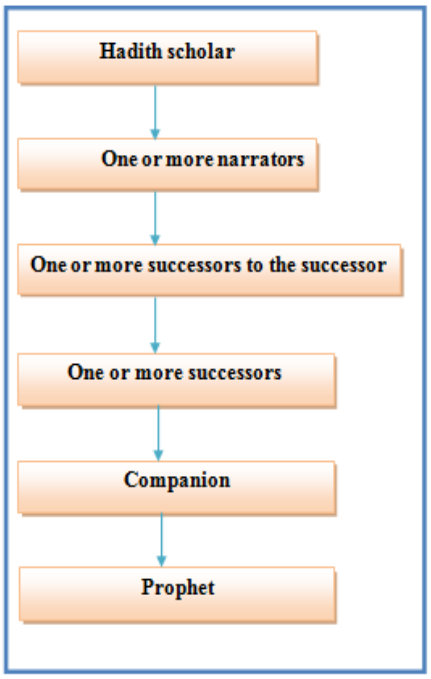

Fig. 2 Composition of Hadith Isnad

2) Depending on the links in the Isnad, i.e. whether the reporter chain is broken or not. The Hadith can be supported, continued, broken, suspended perplexed or hurried.
3) According to the number of reporters involved in each phase of Isnad, In this case, Hadith can be authentic or isolated. The latter being divided into rare, strong or famous Hadith.
4) Depending on the nature of the Text and Isnad, for example, adding a reliable reporter, or opposition from a lower authority to a more reliable person, called irregular. In some cases, a text containing a vulgar expression, an unreasonable remark or an obviously erroneous assertion is rejected by Hadith scholars without regard to the Isnad: such a hadith is called denounced. If an expression or statement is proven to be an addition to the text, it is declared as interpolated.
5) According to the reliability and memory of the narrators; the final judgment on a hadith depends crucially on this factor: verdicts such as sound, good, weak or invented are mainly based on the nature of the narrators in the chain of Isnad [10].

Hadith scholars have established five authentication criteria of the sentences attributed to the Prophet. Three criteria concern the study of the Isnad (1 to 3) and two criteria are devoted to the study of the content (4 and 5). These conditions were invented by 'Imam Al-Bukhari' but they received the approval of all subsequent traditionalists [11]. The five criteria are:

- The righteousness of narrators
- The reliability of narrators' memory
- Continuity between transmitters
- The absence of irregularity
- The absence of hidden defects

For Hadith specialists, "authentic Hadith" is any Hadith which "meets the 5 previous conditions", and there is no doubt on the chain of its narrators.

The study of the Isnad (chain of transmission or reporters) has aroused the interest of many great Muslim scholars to the point that they have become excellent experts in the knowledge of Hadith narrators. It includes the identification of the name of each narrator (reporter), his character (his veracity, his piety, his behavior in public and in private, etc.), his ability and his reputation in matters of memorization and the types of narratives known for: whether they are genuine, weak, made up, etc. In addition, each narrator must be identified by a rating given by other narrators who knew him [12]. So all of these and many other disciplines must be considered to know the extent to which hadith can be used as a basis for belief, for jurisprudence or simply as a point of interest (not attributed to the sayings of the Prophet).

Criticism consists in verifying the reliability of the narrators of a chain of guarantors. The approach aims to decide on the morality, the degree of memorization and the continuity of the transmission between each reporter and his student, in order to verify that each link in the chain is trustworthy.

In examining the narrators of a hadith, genuine or denigrating remarks from recognized experts have proven to be very useful. Table 1 shows an examples of such remarks in descending order of authentication:

Table 1: Hierarchy of narrators

Qualification	Transliteration	Arabic terminology
Prince of believers in Hadith	aminin'Mu-al A	أمير المؤمنين في الحديث
Ruler	Al-Hakim	الحاكم
Master in hadith	Imâm Hûjja	إمام حجة
Great memorizer	Imâm Hafiz	إمام حافظ
Very reliable: excellent memory	Thiaqa, Mutqin	ثقة – ثبت – متقن للحديث
Reliable, good memory	Sadduq, La Baassa Bihi	صدوق – لا بأس به
Good in hadith	Saleh	صالح
Flexible in hadith	Layen al-Hadith	لين الحديث
Not very strong in memory	Layssa Bel-Qawi	ليس بالقوي
Weak	Dha'if	ضعيف
Abandoned	Mutruqal-Hadith	متروك الحديث
Liar	Kadhab	كذاب

The golden rule of Hadith scholars is as follows: the authenticity of a hadith must be based on an impeccable chain of transmission according to the three criteria that we have explained previously (a, b and c) [13]. This task requires the segmentation of Isnad to identify the list of its narrators and transmission tools.

2.2 Segmentation of Arabic text : Approach and previous studies

Given its morphological, syntactic and semantic properties, the Arabic language is considered a difficult language to master in the field of NLP. The first research works, started around the seventies, were concerned with Arabic lexicons and morphology. With the advent of the Internet and search engines, the amount of Arabic documents available in electronic format has become enormous. As a result, several research works for its automatic processing are beginning to emerge. This work has taken various orientations relating to syntax, semantics, information retrieval, information extraction, machine translation, automatic indexing of documents, etc.[14].

The segmentation of a computerized text is the operation of eliminating the segments of its basic elements which are the characters, into constituent elements of different structural levels: paragraph, sentence, phrase, graphic word, morpheme, etc. Segmentation is a central task for natural language processing. It can indeed improve many applications such as syntactic analysis, extraction of multilingual information, automatic summarization, automatic translation, information retrieval for hadith, extraction and visualization of the chain of narrators from Hadith, determining Hadith validity, etc. It corresponds to the question of choosing the most appropriate tag for each word of a text in a predefined inventory.

In NLP field, the notion of a word in a lexical tagging task does not necessarily correspond to a traditional word due to the blind segmentation of texts without syntactic or semantic information. A traditional word can be divided into several units or morphemes (in the case of amalgams or compound words, for example). On the contrary, several words in sequence can be grouped into a single unit: phrases, compound proper names, compound numbers, compound words, etc. Depending on the definition of the lexical units and/or the application, the descriptions of the classes and the morph syntactic tags can include one or more features such as the syntactic category, the lemma, the gender, the number.

Today there are various tools for morph-syntactic tagging, as well as immense resources of annotated corpora for various processing in many languages. The Treebank projects (http://www.cis.upenn.edu/~treebank/home.html)

are examples of creating large annotated corpora. This also assumes the existence of various definitions of lexical units and tag sets depending on the objective. This also raises the crucial questions of the reusability of these linguistic resources for a growing number of applications, their combined reuse in a multilingual context, and the adaptation of tools to other languages. Multiple projects have emerged in this perspective: the evaluation of tools, the standardization and the representation of morph-syntactic description structures [15].

Natural languages, in terms of their writing system, belong to two different families: languages “with separators” and languages “without separators”. The languages known as "with separators" are those which have segmented writing systems, that is to say writings delimited by spaces and where the words are clearly separated by delimiters (space, punctuation marks, special characters, etc.). To this type of language we oppose the so-called languages “without separators”. They present non-segmented writing systems where the words are not separated by spaces and where the borders of the words are not clear. Japanese, Chinese and Thai are the perfect representatives of this second family of languages.

The Arabic language presents a writing system at the intersection of the two families. It is a writing system that combines segmented writing and non-segmented writing. Indeed, part of the Arabic graphic words corresponds to minimal words separated by delimiters. On the other hand, a large part of Arabic graphic words are composed of a series of agglutinated lexical units that can be analyzed in terms of minimal words and clitics and which must therefore be segmented to arrive at the basic units composing them [16].

There are several levels of analysis at which we can stop to identify the different elements that make up the text and define its boundaries. We can stop at the level of the sentence, at the level of the proposition or at that of the phrase. But we can also arrive at level of the graphic word, the lexical units or to go beyond these to arrive at the basic units composing them: the morphemes [17]. Depending on the aim of the analysis to be undertaken: lexical, morphological or syntactic, we can generally consider three main types of application of segmentation:

Tokenization (or word segmentation): This type of segmentation is also called lexical segmentation. It means the segmentation of a text into words or lexical items (tokens).

Morphological segmentation: It aims to isolate the different constituents of lexical items into distinct, smaller units, called morphemes.

Chunking: This type of segmentation is also called syntactic segmentation. It consists of isolating the different constituents of the text into independent units, superior to words, such as propositions, phrases, etc.

By segmentation, we mean here the syntactic segmentation or chunking which consists in segmenting an Isnad and determine the tags of its components (transmission tool, narrator name, name prefix, name or attribute of Prophet, etc.)

There are two families of methods for automatic part of speech tagging: rule-based methods and stochastic methods, all operating in supervised and unsupervised mode. Rule-based methods use contextual linguistic information provided by experts to assign tags to unknown or ambiguous words. Many taggers use morphological and syntactic information to resolve ambiguity caused by unfamiliar words. Some systems go beyond this information by including rules that take other factors such as punctuation or the use of capital letters. The second methods incorporate frequency or probability into the validation process [18]. An alternative of this approach is to calculate the probability of a given sequence of occurring tags. This approach is based on the n-grams method considering that the best tag for a given word is determined by the probability that it occurs with the previous n tags. The method generally used in a stochastic tagger combines the two approaches using tag sequence probabilities and word frequency measures.

For Modern Standard Arabic many works have been done. The best known are based on Markov models, rule-based symbolic systems or even neural networks [19-22]. The most famous are Aramorph, Sebawi [8], APT tagger and Alkhalil [8] and [14]. These systems provide for each word in the Arabic language: the suffixes, the prefixes, the radical, the canonical form (lemma) as well as other information such as the grammatical gender (feminine, masculine), the number (singular, plural) or the time (present, perfect past,).

In the case of Hadith Isnad, the work of tagging is a new and difficult task because of its morphological and structural aspects. The structure of Isnad is totally different to any other type of Arabic sentence.

Consider the following example of Hadith:

حدثنا قتيبة قال حدثنا الليث عن يزيد بن أبي حبيب عن أبي الخير عن عبد الله بن عمرو أن رجلا سأل رسول الله صلى الله عليه وسلم أي الاسلام خير قال تطعم الطعام وتقرأ السلام على من عرفت ومن لم تعرف

Qaytibah told us Alith from Yazid ibn Abi Habib from Abi Al-Khair from Abdullah bin Amr A man asked the Messenger of Allah (PBUH): ”Which act in Islam is the best?” He (PBUH) replied, ”To give food, and to greet everyone, whether you know or you do not.”

The Isnad is:

حدثنا قتيبة قال حدثنا الليث عن يزيد بن أبي حبيب عن أبي الخير عن عبد الله بن عمرو أن رجلا سأل رسول الله صلى الله عليه وسلم أي الاسلام خير قال

Qaytibah told us Alith from Yazid ibn Abi Habib from Abi *Al-Khair from Abdullah bin Amr A man asked the Messenger of Allah (PBUH): ”Which act in Islam is the best?” He replied*

The part of Isnad tagging attribute the tag ‘transmission tool’ to the word ‘حدثنا’, the tag ‘narrator name’ to noun ‘يزيد بن أبي حبيب’ and so on.
Some attempts to detect Isnad patterns or show the Isnad tree are described with their approach, technics, results and limitations in [23] and [24].
Hidden Markov models (HMMs) are statistical tools based on solid theory. They are characterized by great efficiency and flexibility in the algorithms they offer for modeling the addressed problems. These models have recently been widely used in the field of text segmentation and part of speech tagging for many natural languages [25]. In the next section, we present the theoretical concepts and required tools for the understanding and use of HMM Isnad chunking.

3. The HMM-Based Chunking of Isnad

A stochastic process $\Omega\{X\ t,\ t \in T\}$ is a set of random variables defined on a probability space Ω, often denoted by Xt , t ∈{1,...T}, t represents time.
A discrete Markov chain of order n is a discrete stochastic process X = {Xt|t = 1,...T} with Xt discrete random variables, satisfying the Markov property:

$$p\left(X_t = s_t \mid X_{t-1} = s_{t-1}, \ldots, X_{t-n} = s_{t-n}, X_{t-n-1} = s_{t-n-1}, \ldots, X_1 = s_1\right) = p\left(X_t = s_t \mid X_{t-1} = s_{t-1}, \ldots, X_{t-n} = s_{t-n}\right)$$
$$\forall t \in [1,T]\ et\ s_1, \ldots, s_2 \in S$$

S= {s1…sT} represents the set of states; the current state depends only on the previous n states.

$$p(X_t = s_t \mid X_{t-1} = s_{t-1}, \ldots, X_1 = s_1) = p\left(X_t = s_t \mid X_{t-1} = s_{t-1}\right)$$

For a discrete Markov chain of order 1, only the current state and its predecessor are considered. A Markov chain of order 1 is said to be stationary. i.e. the current state depends only on a previous state and it does not depend on time,
A one-dimensional Hidden Markov Model (HMM) is a stationary Markov chain where the observation is a probabilistic function of the state, which is characterized by a doubly stochastic state system constituting two processes. The first is a process of state change called hidden or internal process and which is not observable. However, it can be observable through a second emission process called the external process. The HMM behavior is therefore linked to the two sequences of random variables which are associated respectively with the observable and hidden component of the stochastic model [25].

The hidden sequence corresponds to the sequence of states q1, q2… qT represented by Q (1: T) with qi ∈{ S1 ,...,SN }the set of n model states.
The observable suite corresponding to the sequence of observations O1, O2,…,OT denoted by Q (1: T).
In our case, the part of Isnad tagging is assumed to be a Markovian process with unknown parameters. It treats the text of Isnad as a sequence of hidden states (the sequence of tags), each state producing emissions (the observable phrase of the Isnad). In the Markov model, we consider that the choice of a tag for a phrase must depend on the n previous tags, and of the word itself. Moving from one state to another, from one tag to the next, is called a transition. Algorithm such as Viterbi [26] make it possible to identify the sequence of labels maximizing the probability of transitions for a sequence of given words. To estimate the probability that tag X is followed by tag Y, we can base on the frequencies observed in a training corpus, with the following formula:
P(x|y) = freq(x , y) / freq(y)

The input of the HMM-Based chunking of Isnad is an Isnad (chain of narrators) or a set of numbered Isnad stocked in an XML file (Fig.3).

Fig. 3 A part of XML File

The XML file is considered as a database containing in each element the number of Hadith as in "Sahih Alboukhari", the Matn (text of Hadith), the Isnad, the book, the part (Albab), the companion and the final speaker.
Fig. 4 presents the architecture of the proposed method to segment Isnad and determine the tag of each part.

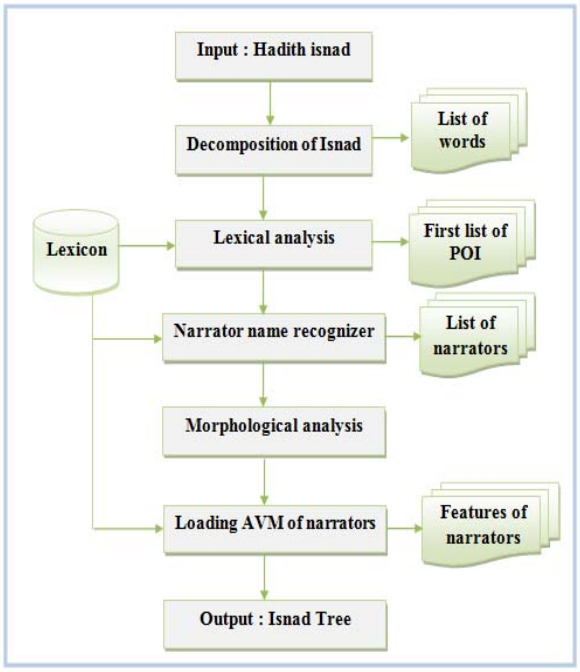

Fig. 4 System architecture

To build our POI tagger (chunker) for Hadith Isnad, we have implemented a lexical analysis module which tries in a first step to attribute for each word obtained after the decomposition phase its proper tag. In general this step allows us to determine the transmission tools (قال، حدثنا، عن، أخبرنا، أن...). All the possible transmission tools are stocked in the lexicon. This lexicon is a set of XML documents. The next main module is the morphological module. Its main objective is to find the tag of each part of isnad using HMM model. This module is based on a narrator name recognizer witch permit to collect a set of words as a narrator name [9].

In the HMM model, an "emission probability" is the probability of observing the input isnad or sequence of words W given the state sequence T, that is P(W|T). Also, from the state transition probabilities we can calculate the probability P(T) of forming or following the state sequence T. Let's consider the HMM model (Table2, 3 and 4) of the Hadith Isnad examples:

Table 2: POS tagging of Isnad **:** حدثنا مالك عن نافع عن بن عمر

POI	بن عمر	عن	نافع	عن	مالك	حدثنا
Tag	Narrator	T. Tool	Narrotor	T. tool	Narrator	T. tool
Arabic Tag	راو	أداة إخبار	راو	أداة إخبار	راو	أداة إخبار

Table 3 : POS tagging OF ISNAD : روى الزهري عن نافع عن أبيه

POI	أبيه	عن	نافع	عن	الزهري
tag	Prefix name	T. Tool	Narrator	T. tool	Narrator
Arabic Tag	استعاضة	أداة إخبار	راو	أداة إخبار	راو

Table 4 : POS tagging OF ISNAD : **عن نافع قال أخبرنا سعيد عن أبي هريرة**

POI	أبي هريرة	عن	سعيد	أخبرنا	قال	نافع
Tag	Narrator	T. tool	Narrator	T. tool	T. tool	Narrator
Arabic Tag	راو	أداة إخبار	راو	أداة إخبار	أداة إخبار	راو

If we consider a single Isnad, we obtain a table that contains each part of Isnad and its proper tag.

4. Implementation and Discussion

4.1 The training corpus

We have selected a training corpus of 100 Isnad from "Sahih Alboukhari". The entrees of the corpus are chosen so that they cover the majority of isnad structures. The selection of the corpus is based firstly on an in-depth study of the Isnad structures and on the other hand, based on discussions with experts in this field. Each Isnad is manually tagged with the possible tags (name of narrator, transmission tool, konia, etc.). We used 13 tags described in [2] and [5]. Fig .5 shows a part of the tagged training corpus.

Fig. 5 Part of the training corpus

4.2 Learning and Results

A trigram language model was built for the tagged training corpus. The trigram language model computes lexical probabilities. Then, we obtained the POI tag sequences from the training corpus and created a trigram Arabic language model based on the POI tag corpus. The benefit of this tag model is that it allows calculating the probability of one tag following another tag (contextual model) [19]. Since we use trigram models it is possible that some trigrams were never observed in the training corpus. The probability of unseen trigrams cannot be assigned to zero because this will cause the (computed) probability of an entire observed sequence to be zero. We rather use the back-off smoothing technique [25] so that the model backs off to a bigram model. Similarly, if the bigram was never observed during training, then we would back off to a unigram model. For the purposes of this smoothing, we have created unigram, bigram and trigram lexical and contextual language models. Next, lexical and contextual probabilities were used to build the HMM model's parameters as follows : contextual probabilities were stored in parameter A as transitions probabilities and lexical probabilities were stored in parameter B as the emission probabilities. Once matrices A and B are computed, search needs to be performed to find the POS tag sequence that maximizes the product of the lexical and contextual probabilities. The Viterbi algorithm [26] is used for a faster computation of the optimal path.

So, the experimental work was carried out in three main stages:

1) Step of defining the set of tags and building a training corpus.

2) Step of estimating the parameters of HMM.

3) Step of automatic tagging and re-estimation of the parameters of HMM.

In order to carry out these last two steps, we have developed a java application, consisting of two main modules, a learning module and an automatic labeling module that automatically labels the entry isnad, which is manually corrected for re-estimation of the parameters of the hidden Markov model. Fig. 6 shows the graphical user interface used to enter and decompose the Isnad and presents the results of applied method used to tag the input Isnad.

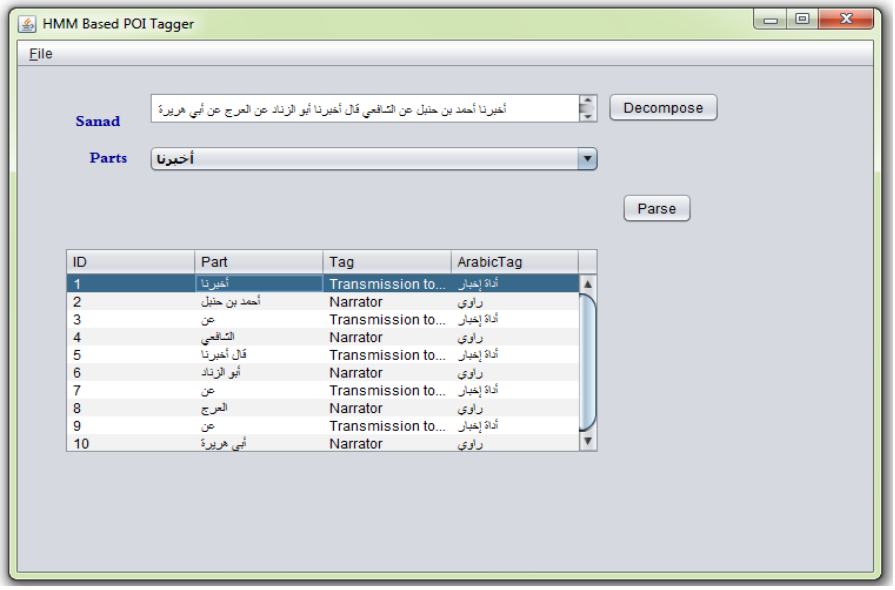

Fig. 6 Result of the HMM Based POI Tagger

We also can find and analyze several Asanid for one Hadith. The system trace the parse trees and highlight the common narrators as shown in Fig .7.

Fig. 7 Asanid Tree for a Hadith

5. Conclusion

In this paper, we have proposed a statistical method based on HMM to segment Hadith Isnad and determine the appropriate tag of each component. To do this, we organized our work according to three main steps. First, we collected and prepared all the necessary linguistic data: information about the narrators, corpus of work, lexicon, possible tags and corpus of learning. Then, we developed an Isnad segmentation method, based on this corpus. Finally, we ended with a quantitative and qualitative evaluation of our prototype. We used Java programming language and XML technology to ensure the interoperability and the reusability of NLP tools serving the Matn and Sanad criticisms in classification and Evaluation of Hadith. The prototype was tested on 2661 Isnad collected from "Sahih Alboukhari". 2489 Isnad have been fully and correctly segmented and labeled. The obtained result achieved F-scores of 93.53%.

References

[1] M. Najeeb, A. Abdelkader, and M. Al-Zghoul, ‘‘Arabic natural language processing laboratory serving Islamic sciences,’’ Int. J. Adv. Comput. Sci. Applic., vol. 5, no. 3, pp. 114–117, 2014.

[2] M. Najeeb, A. Abdelkader, M. Al-Zghoul, A. Osman «A Lexicon for Hadith Science Based on a Corpus» (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) ,, 1336-1340. 2015.

[3] A. Osman, M. Najeeb, A. Abdelkader, and M. Al-Zghoul, “Hadith Graduation as a Service on cloud computing,” International Conference on Cloud Computing ICCC 2015. The College Of Computer and Information Sciences at Princess Nourah bint Abdulrahman University, Riyadh, Kingdom of Saudi Arabia on April 27-28, 2015.

[4] M. Najeeb, “Towards Innovative System for Hadith Isnad Processing,” International Journal of Computer Trends and Technology (IJCTT) V18(6Dec 2014. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group, pp. 257-259, 2014.

[5] M. Najeeb,, “XML database for hadith and narrators,” American Journal of Applied Sciences 13, 1, pp. 55-63. 2016.

[6] M. Najeeb, “Multi-agent system for hadith processing,” International Journal of Software Engineering and Its Applications 9, 9, pp. 153-166, 2015.

[7] A. Abdelkader, M. Najeeb, M. Alnamari and H. Malik. “Creation of Arabic Ontology for Hadith Science,” International Journal of Advanced Trends in Computer Science and Engineering. IJATCSE Volume 8, No.6, pp. 3269-3276, 2019.

[8] A. Abdelkader, M. Najeeb, M. Alnamari and H. Malik. “How Can Existing NLP Tools of Arabic Language Serve Hadith Processing,” International Journal of Computer Engineering and Technology (IJCET) Volume 10, Issue 06, pp. 22-31, 2019.

[9] M. M. Al-Azami, Studies in Hadith Methodology and Literature. Indianapolis, IN, USA: American Trust, 1978.

[10] Islam Web. Accessed: Feb. 05, 2022. [Online]. Available: http://www. islamweb.net

[11] Dorar. Accessed: Jan. 28, 2022. [Online]. Available: http://www.dorar.net

[12] إسماعيل رضوان، طالب أبوشعر، "منهج الحكم على الأسانيد"، مكتبة وطبعة دار المنارة، 2006

[13] K. Faidi, R. Ayed, I. Bounhas, and B. Elayeb, ‘‘Comparing Arabic NLP tools for Hadith classification,’’ Int. J. Islamic Appl. Comput. Sci. Technol., vol. 3, no. 3, pp. 1–12, 2015

[14] I. Bounhas, “On the Usage of a Classical Arabic Corpus as a Language Resource: Related Research and Key

Challenges," Published in ACM Trans. Asian & Low, DOI:10.1145/3277591, 2019.

[15] A. Abdelkarim, D. Boumiza and R. Braham, "A categorization algorithm for the Arabic language," International Conference on Communication, Computer and Power (ICCCP'09), Muscat, February 2009.

[16] A. Azmi, A. Al-Qabbany and A. Hussain, "Computational and natural language processing based studies of hadith literature: A survey," Artif Intell Rev manuscript, 2019.

[17] E. Brill, "Some Advances in Transformation Based Part of Speech Tagging," In proc. Of ICAI'94 (The Twelfth International Conference on Artificial Intelligence) 722-727, 1994.

[18] S. Köprü, "An efficient part-of-speech Tagger for Arabic," Proceedings of the 12th international conference on Computational linguistics and intelligent text processing (CICLing'11), Tokyo, Japan, 2011.

[19] K. Duh and K. Kirchhoff, "POS Tagging of Dialectal Arabic: A Minimally Supervised Approach," In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, Michigan. Association for Computational Linguistics,pp. 55–62, 2005.

[20] T. Brants, "statistical part of speech tagger," In proc. of ANLP'2000 (the 6th Conference on Applied Natural Language Processing) : 224-231, 2000.

[21] M. Diab., H. Kadri. and J. Daniel, "Automatic Tagging of Arabic Text : From Raw Text to Base Phrase Chunks," In proc. of HLTNAACL'04 (Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics), pp. 149–152, 2004.

[22] S. Khoja, "APT : Arabic Part-of-speech Tagger," In proc. of NAACL'2001 (the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics) : 20-26, 2001.

[23] J. Paddy, "Scientific efforts to serve the Sunnah on the internet websites display, analyze and evaluate," in Siminar of efforts in the Sunnah Service, Sharjah University.

[24] M. Najeeb, "A Novel Hadith Processing Approach Based on Genetic Algorithms," IEEE Access, Vol 8, 2020.

[25] M. Albared, N. Omar, M. AbAziz, "Developing a Competitive HMM Arabic POS Tagger Using Small Training Corpora," In: N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 288–296, 2011.

[26] G. D. Forney, "The Viterbi Algorithm," In proc. of the IEEE Transactions on Information Theory, pp. 263-278, 1973.